REGRESSION LINEAIRE

Le but d'une régression linéaire est de trouver **la meilleure relation affine entre deux séries de données**.

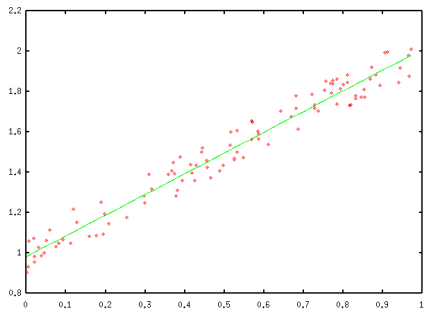
I. VERIFICATION D'UNE RELATION GRACE A UNE REGRESSION LINEAIRE

On possède deux séries de données entre lesquelles on veut trouver ou vérifier une relation.
Il est très facile de vérifier qu'une relation est linéaire (ou affine) entre deux séries : il suffit de vérifier si les points sont alignés. En revanche, il est plus difficile de montrer qu'une relation est logarithmique, parabolique…
L'idée est donc de construire, à partir des deux séries d'origine, deux séries de données entre lesquelles il existe une relation affine. On appelle cette opération **« linéariser une relation ».**

Exemple : On veut vérifier qu'une constante de vitesse k vérifie la relation d'Arrhenius k=Aexp(-Ea/RT) et on dispose pour cela de valeurs de k à différentes températures T.
Il faut donc trouver une relation affine entre deux séries de données bien choisies en linéarisant la relation d'Arrhenius : $k = Aexp\left(-\frac{Ea}{RT}\right) \iff lnk = lnA - \left(\frac{Ea}{R}\right) \cdot \frac{1}{T}$.
On trouve une relation du type y=ax+b avec y=lnk ; a=-E_a/R ; x=1/T ; b=lnA.
Conclusion : Si la relation d'Arrhenius est verifiée, on a alors une relation affine entre lnk et 1/T.

II. PRINCIPE DE LA REGRESSION LINEAIRE

Si on place dans un plan les points correspondants aux couples formés par les deux séries de données fournies, on obtient ce qu'on appelle un nuage de points. **Effectuer une régression linéaire entre les deux séries consiste à trouver la droite qui passe au plus près de l'ensemble de ces points.**



On peut alors déterminer trois paramètres :

- **la pente de la droite a, et son incertitude.**
- **son ordonnée à l'origine b**
- **les coefficients de corrélation (ou de régression), par exemple r^2 (compris entre 0 et 1). Les coefficients de corrélation permettent d'évaluer la proximité de la droite par rapport à la série de données.**

C'est la calculatrice qui effectue cette opération.

III. COEFFICIENTS DE CORRELATION r ET r^2 : QUALITE D'UNE REGRESSION LINEAIRE

La calculatrice est une machine : elle n'émet pas de jugement sur les opérations qu'on lui fait faire : si une régression linéaire est demandée, la calculatrice proposera l'équation d'une droite (même si les points de l'échantillon ne sont pas du tout alignés…). Il faut donc évaluer la qualité de la régression linéaire.
Pour ce faire, on utilise deux critères :

- On effectue une vérification visuelle : on **trace le nuage de points et on vérifie qu'ils semblent alignés**.
- On utilise les coefficients de corrélation **r et r^2** :

Plus la valeur de r^2 est proche de 1, meilleure est la qualité de la régression linéaire.
Usuellement, on considérera qu'une régression linéaire est satisfaisante si r^2>0,99*.

Remarque : Donner suffisamment de chiffres significatifs pour r^2 pour montrer qu'il est supérieur à 0,99.
**Attention, se méfier du seul coefficient de corrélation, faire aussi une vérification visuelle.*

IV. MISE EN ŒUVRE

Méthode : Pour vérifier une relation grâce à une régression linéaire, il faut :

- **« Intuiter » la relation entre les deux séries de donner, et la linéariser.**
- **Calculer si nécessaire les nouvelles séries de données entre lesquelles la relation est supposée affine.**
- **Tracer le nuage de points et effectuer la régression linéaire à la calculatrice (ou à l'aide d'un logiciel)**
- **Vérifier si cette régression est de bonne qualité grâce au tracé (vérifier visuellement que les points semblent alignés, qu'ils sont proche de la droite et répartis aléatoirement au dessus en en dessous de la droite) et au coefficient de corrélation (vérifier que $r^2>0,99$). Si c'est le cas, on pourra considérer que la relation est effectivement correcte. Sinon, c'est que l'hypothèse de départ est à revoir.**

Chacun doit savoir effectuer une régression linéaire avec sa calculatrice, et est responsable de la maitrise de son matériel. Quelques informations sont présentées ci-dessous, bien sûr non exhaustives. En cas de besoin, lisez votre mode d'emploi.

- **Casio :** Aller dans le menu [STAT] et rentrer deux séries de données entre lesquelles la relation est supposée affine, chacune dans une liste. Visualiser le nuage de point en tapant [GRAPH] puis en choisissant un graphe (GPH1 par exemple). Apparaît alors sous le graphe un menu : pour obtenir une régression linéaire y=ax+b, taper [X]. Une fenêtre LinearReg donne alors les valeurs a, b et r^2 en précisant l'équation de la droite.

Attention : si plusieurs listes ont été créées, il faut préciser quelles listes jouent les rôles de « x » et « y ». Pour cela, appuyer sur [GRAPH] et choisir [SET].

Pour visualiser la droite de régression linéaire, choisir [DRAW] en bas à droite de la fenêtre LinearReg.

- **TI :** Ouvrir l'éditeur de données : [STAT][EDIT][1 : Edit…]. Rentrer deux séries de données entre lesquelles la relation est supposée affine, chacune dans une liste.

Pour visualiser le nuage de point: ouvrir l'éditeur de graphes avec [2ND][STAT PLOT][1]. Cocher [On] puis choisir le type de tracé, par exemple le nuage (tracé discontinu). Entrer le noms des listes qui jouent les rôles de « x » et « y ». Choisir le type de marque (par exemple le carré). Appuyer ensuite sur la touche GRAPH (les points doivent a priori semblés alignés…). Pour ajuster la fenêtre, régler le zoom : [ZOOM][9 : ZoomStat].

Pour effectuer la régression linéaire, entrer ensuite dans [STAT][TESTS] et choisir la régression linéaire [LinRegTTest]. Indiquer la liste choisie pour x, et celle choisie pour y, Aller sur [Calculate]. La calculatrice donne la pente, l'ordonnée à l'origine et le coefficient de corrélation r^2.

Pour visualiser la droite de régression linéaire, appuyer sur la touche [Y=] puis [CLEAR]. Entrer ensuite dans [VARS][5 : Statistics…] et sélectionner [EQ][1 : RegEQ]. Visualiser ensuite le graphe en appuyant sur [GRAPH].

V. INTERVALLE DE CONFIANCE

Certains logiciels fournissent l'incertitude sur la pente et l'ordonnée à l'origine ainsi obtenues.

Par exemple, sur TI : choisir la régression linéaire [LinRegTTest]. Indiquer la liste choisie pour x, et celle choisie pour y, ainsi que l'intervalle de confiance souhaité pour la pente (par exemple 0,99). Aller sur [Calculate]. La calculatrice donne la pente avec son intervalle de confiance.

D'autres logiciels fournissent l'incertitude type u(p) sur la pente. Il faut alors multiplier cette incertitude type par le coefficient de Student t correspondant au nombre de valeurs de la liste et à l'intervalle de confiance souhaité. Si on ne dispose pas de table, prendre t=2. On a alors « pente = valeur donnée ± 2.u(p) ».

En pratique, les calculatrices graphiques et les logiciels avec tableur permettent le tracé du graphe y(x) et donnent les coefficients a et b et le coefficient de corrélation r^2. Certains tableurs donnent en plus les incertitudes-types sur a et b (par exemple Régressi et Latis Pro). En fonction du logiciel utilisé, on cherchera dans la notice la manière d'obtenir ces données (si possible).

VI. QUELQUES CONSIGNES

- Attention à l'erreur classique qui consiste à échanger les deux listes (la régression ne s'effectuant pas dans le bon sens).
- TOUJOURS tracer la droite et donner la valeur de r^2 pour justifier le fait que la régression est valable. Ne pas donner toutefois trop de chiffres significatifs (trois suffisent).
- Indiquez toujours ce que vous tracez. (Par exemple : « on trace lnk en fonction de 1/T »)

EXERCICES D'APPLICATION : REGRESSIONS LINEAIRES

Exercice 1 : Loi de vitesse

On étudie le mouvement uniformément accéléré d'un mobile, sa vitesse est alors donnée par la loi $v=at+v_0$. On obtient les données suivantes :

t(s)	1	2	3	4	5
v($m.s^{-1}$)	1,0	1,6	2,2	2,8	3,4

1. La loi semble-t-elle vérifiée ?
2. Déterminer l'accélération a du mobile et sa vitesse v_0.

Exercice 2 : Cinétique chimique

La concentration C d'une espèce chimique est mesurée au cours du temps. On obtient les données suivantes :

t(s)	20	40	60	80	100	120
C($\mu mol.L^{-1}$)	278	192	147	119	100	86

1. Réaliser les régressions linéaires suivantes, en donnant l'équation et le coefficient de corrélation :
   - C en fonction de t
   - ln(C) en fonction de t
   - 1/C en fonction de t
2. Avec quelle loi les résultats expérimentaux s'accordent-ils le mieux ?

Exercice 3 : Raideur d'un ressort

L'énergie E d'un ressort dépend de son allongement x selon la loi : $E=kx^2$.

On obtient les données suivantes :

x(cm)	10	-9,3	8,6	-7,9	7,2	-6,5	5,8
E(mJ)	100	86,5	74,0	62,4	51,8	42,2	33,6

1. La loi est-elle vérifiée ?
2. Si oui, quelle est la valeur de la constante de raideur k du ressort ?

Réponses

Exercice 1 :

1. *Droite d'équation y = 0,6x+0,4 ; r^2=1. Les points semblent alignés, bon coefficient de corrélation : la loi semble vérifiée.*
2. *$a = 0,6\ m.s^{-2}$ et $v_0 = 0,4\ m.s^{-1}$.*

Exercice 2 :

1. *Avec C en $mol.L^{-1}$ et t en secondes :*
   $C = f(t) : C = -2.10^{-6}.t + 0,0003$; r^2=0,89
   $ln(C) = f(t) : lnC = -0,0115.t – 8,06$; r^2=0,97
   $1/C = f(t) : 1/C = 80,2.t + 1990$; r^2=1
2. *Coefficient de corrélation maximal, meilleur alignement : la loi suivie semble être une relation affine entre 1/C et t. Pente : 80,2 $mol^{-1}.L.s^{-1}$ (Intervalle de confiance à 95% donné par la calculatrice : [80,0 ; 81,0]).*

Exercice 3 :

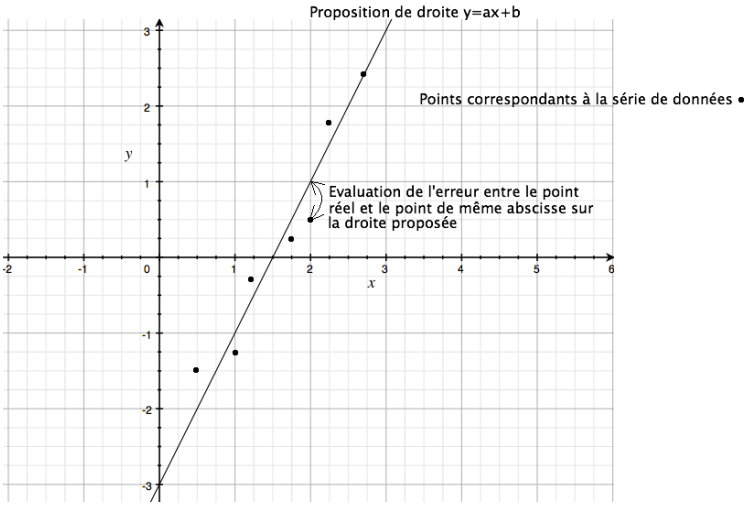
1. *E(en J) en fonction de x^2 (avec x en m) est bien une droite d'équation $E=10,01.x^2 – 8.10^{-5}$ (ordonnée à l'origine négligeable, r^2=1) : la loi est vérifiée.*
3. *$k=9,99.10^{-2}\ J.m^{-2}$ ($=N.m^{-1}$) (Intervalle de confiance à 95% donné par la calculatrice : [0,00998 ; 0,0100]).*

Pour information :

- ***Comment la calculatrice choisit-elle la meilleure droite ?***

Il y a de nombreux critères existant, le plus courant étant celui des « moindres carrés ».

Pour cela, la calculatrice propose une droite d'équation y=ax+b, et évalue l'erreur commise entre le point réel pour chaque couple (x,y) des séries de données et le point d'abscisse x de la droite proposée.


Proposition de droite y=ax+b
Points correspondants à la série de données
Evaluation de l'erreur entre le point réel et le point de même abscisse sur la droite proposée
x
y

Puisque l'erreur commise est tantôt positive (point au dessus de la droite) et tantôt négative (point en dessous de la droite), la somme de ces erreurs sera quasiment nulle même si les points sont très éloignés de la droite. Cette somme est donc un mauvais critère d'évaluation de la qualité de la régression linéaire. On choisit alors comme grandeur d'évaluation la somme des carrés des erreurs : plus cette somme est faible, plus les points sont proches de la droite.

$$\sum_{i} r_i^2 = \sum_{\text{mesures}} \left[Y_{i\ \text{mesuré}} - \left(aX_{i\ \text{mesuré}} + b \right) \right]^2$$

La calculatrice cherche à minimiser cette grandeur, d'où l'appellation de « critère des moindres carrés ».

- ***Méthode de régression linéaire et incertitude…***

Cette méthode n'est valable que si :

- *l'incertitude sur les mesures des valeurs en abscisse est négligeable*
- *l'incertitude sur les mesures des valeurs en abscisse est constante*
- *Il n'y a pas d'erreur systématique*

Si ça n'est pas le cas, il faut utiliser la méthode dite « du χ^2 » qui consiste à tracer des ellipses correspondant aux incertitudes sur chaque point, et de tracer la droite passant au plus près de chaque ellipse. Ainsi, la droite passe près des points les plus certains et plus loin des points plus incertains.